基于DTM模型的国内外政府数据研究主题挖掘及比较分析

马海群[1] 孙正轩[2]

黑龙江大学信息管理学院 哈尔滨 150000

摘要：[目的/意义]对国内外政府数据的文献进行研究，分析国内外研究主题和演化趋势的差异，对于了解国内外政府数据研究现状，推动国内政府数据研究的创新发展具有重要的参考意义。[方法/过程]选取2010年至今政府数据的相关文献，基于动态主题模型（Dynamic Topic Model,DTM）对国内外政府数据文献进行分析，通过构建高频词词云图、分析主题强度演化趋势及主题关键词演化路径三个方面对比国内外政府数据研究的现状及差异。[结果/结论]国内外的政府数据相关研究存在较大差异：①国内的研究主题数量较多，研究内容比较细化，而国外的研究主题数量较少，主要对政府数据的质量、框架模型和评估方法进行研究。②国内的研究主题强度整体较低，起伏波动较大，国外的研究主题强度整体较高，主题演化趋势中较为稳定，波动较小，具有较高的传承性。③国外侧重“政府数据评估”的相关研究，主题强度最高，国内关注度较低，虽然在最近有所提升，但与国外相比关注度仍然较低。

关键词：政府数据；DTM模型；主题挖掘；比较分析

分类号：G250

1 引言

政府数据是政府使用信息技术主动向公众免费开放的数据信息，主要特征包括开放性、授权性和机器适用性，在数字时代具有很高的研究价值[1]。随着信息技术和大数据产业的蓬勃发展，数据资源成爆发式增长，信息技术和智能技术得到了广泛的应用，各种数据正在以其独特的方式彰显自身的价值，其中政府数据在社会公共事物的治理中发挥着举足轻重的作用。各国政府步入数字时代的过程中，积累了大量的政府数据，这些数据是非常重要的非物质财富，记录着政府以往进行的数字活动[2]。政府数据的合理使用可以为人们的工作和生活带来极大的便利条件，因此政府需要加强对数据的治理以保证数据的质量，实现数据开放、

[1]马海群（1964-），男，mahaiqun@126.com，教授，博士，博士生导师，研究方向：信息政策与法律
[2]孙正轩（1998-），男，通讯作者，2022级硕士研究生,研究方向：信息政策与法律

数据共享、促进数字经济的稳定发展和智能社会的平稳运行。

本文选用中国知网和 Web of Science 两大平台中的政府数据相关文献，采用主题挖掘的方法，通过挖掘国内外政府数据相关文章中的研究主题，分析国内外政府数据的研究现状、根据主题强度的变化趋势以及主题自身的演化状况，揭示国内外相关研究的异同点，以期为我国未来政府数据的发展提供新的研究思路和参考价值。

2 相关研究

2.1 政府数据研究

政府数据是国家最重要的战略资产，在大数据背景下其最大作用应当是对民众进行开放，与民众共享，推动政府服务效率的同时刺激国家的经济增长。因此，针对政府数据如何有效开放、怎样搭建高效率的政府开放平台、开放过程中隐私的保护和数据的安全问题，成为了广大学者关注的焦点。马海群等[3]运用 WSR（WuH-Shili-Renli System Approach）方法论对美国和加拿大政策框架进行分析和评估，并且对政策的执行力影响因素进行了深入的研究。黄如花等[4]根据我国政府数据资源管理政策的文本进行分析，认为我国政府数据开放政策关注数据获取与利用涉及较多，对于数据生产、归档和保存、数据监管等问题的关注度仍有待提高。陈美[5]从隐私风险控制的角度对政府数据进行研究，分析了国内外政府数据开放风险的研究动态，指出我国理论研究仍明显滞后，取法个人隐私保护的法律法规。翟军等[6-8]对政府数据的质量进行评价，以元数据、数据的标准和检索等内容进行深入研究。马海群等[9-11]采用基于结构方程法、神经网络算法和 DEA 分析法对数据开放网站的质量和效率进行评价和研究。

纵观我国对于政府数据相关的研究，国内学者主要围绕着政府数据开放政策法规、政府开放数据管理、政府开放数据平台研究、政府开放数据利用等方面进行研究，少有运用主题模型对国内外政府数据的研究热点和演化趋势进行系统性的概括和总结。

2.2 DTM 主题模型研究

主题模型（Top Model）是一种以非监督学习的方式对文本及的隐含语义结构进行聚类的统计模型[12]。自从 2003 年 LDA 主题模型由 Blei 首先提出后（Blei, Ng, & Jordan, 2003），便掀起了主题识别与挖掘的热潮[13]。多数学者从论文、

政策文本等资料中，识别并挖掘主题，从而掌握该领域的研究现状。但是 LDA 模型仅能挖掘主题静态特征，无法分析不同主题在时间序列中的演化趋势，于是 2006 年 Blei 在 LDA 模型的基础上提出了动态主题模型（Dynamic Topic Model,DTM）[14]。DTM 模型弥补了 LDA 模型只能对静态文本建模的缺陷，已经发布便被广泛使用。邱均平等[15]运用 DTM 模型对国内外话语权研究的主题进行对比分析，以破解我国话语权困境，指引未来话语权研究的发展方向。吴瑞鹏等[16]通过量化分析 60 份美国人工智能战略报告，发现美国的人工智能战略中心的转变，并指出了中国面临的挑战和发展建议。章成志等[17]利用 DTM 模型和社区发现技术的话题演变方法，改善了传统话题模型的表现，解决了无法确定与结果可解释性差的问题。

综上所述，虽然 DTM 模型已经被应用在话语权、社区发现技术和战略分析等相关领域，但是在政府数据领域运用 DTM 模型研究主题演进和演化趋势的研究成果较为欠缺，鉴于此，本研究基于 DTM 模型挖掘了政府数据 2010 年至今的主体演化趋势和变化规律，分析出不同时间阶段的热点主题和关键问题，剖析政府数据研究的演化路径，为进一步推动我国政府数据的研究发展提供参考价值。

3 研究方法

本文使用动态主题模型对 2010 年至今国内外政府数据相关文献进行研究，具体研究如下。

3.1 动态主题模型

动态主题模型（Dynamic Topic Model,DTM）是对潜在狄利克雷分配（Latent Dirichlet Allocation,LDA）扩展的一种无监督机器学习的生成式模型，由文档集层、主题层和特征词层 3 层结构组成，可以将文本分成多个时间切片，并且假设每个时间切片中的文档分布都服从相同维度的主题模型，但是每个时间切片中的主题分布和词分布都是随时间变化的。DTM 根据前一个时间片中的主题和词分布来预测后一个时间片中的主题和词分布,从而形成一个动态的主题演化过程，具体原理如图 2 所示。

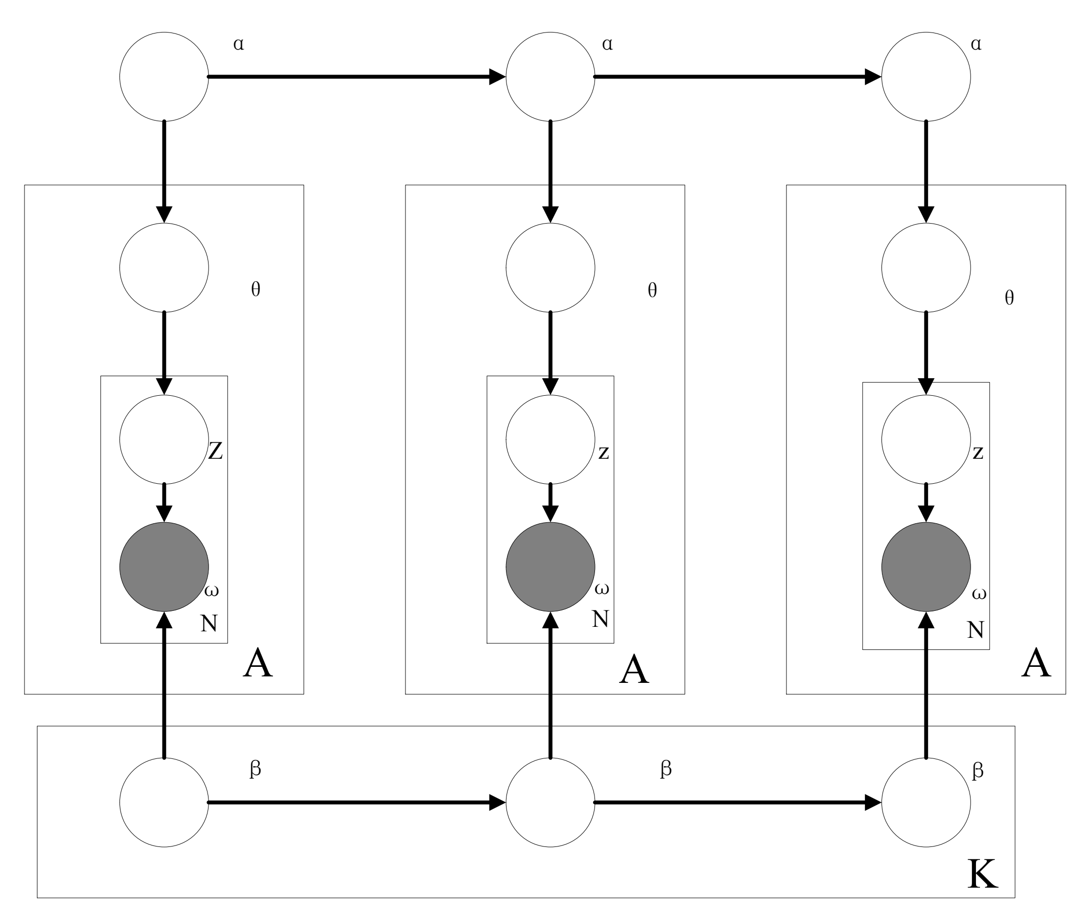


图 2 DTM 模型原理

图 2 中 K 表示主题个数,A 表示时间片内的文档数量,N 表示文档中词的数量,z 和ω表示最终生成的主题和主题词,α与β表示狄利克雷先验分布参数,α是每个文档可能的主题分布,β是每个主题可能的词分布, Θ是该文档的主题模型,表示服从带有参数α的狄利克雷分布。主题 z 由多项分布 Θ选取,与主题相关的词ω由 z 和β的多项分布选取。在每个时间片段 t 下,文档主题分布α_t以及与主题相关的词分布$\beta_{t,k}$均依赖于上一时间片段中的α_{t-1}、$\beta_{t-1,k}$,上一个时间段 t-1 生成这个时间段 t 的$\beta_{t,k}$。通过迭代生成过程,每个时间片段中的文档主题分布和词分布都会受到上一个时间片段的分布参数的影响,从而实现文档主题的连续演化，见公式（1）和（2）：

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$$ 公式（1）

$$\beta_{t,k}|\beta_{t-1,l} \sim N(\beta_{t-1}, \sigma^2 I)$$ 公式

（2）

DTM 模型在 t 时间片段内连续文档主题的生成过程如下：

1）生成主题词分布：$\beta_{t,k}|\beta_{t-1,l} \sim N(\beta_{t-1}, \sigma^2 I)$。

2）生成主题分布：$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$。

3）对于每个文档（文献）：

a. 生成 $\eta \sim Mult(\alpha_t, a^2 I)$

b. 对于每个词：

i. 生成 $Z \sim Mult(\pi(\eta))$；

ii. 生成 $w_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$

其中，函数 π 是多项式到正态分布的映射，见公式（3）

$$\pi(\beta_{t,k})_{\omega} = \frac{\exp(\beta_{t,k,\omega})}{\sum_{\omega} \exp(\beta_{t,k,\omega})}$$ 公式（3）

3.2 研究设计

本文研究路线图如图 1 所示，主要研究步骤如下：（1）从 CNKI 和 Web of Science 上获取国内外政府数据开放研究的相关文献。（2）对文本进行分词，构建停用词表，文本向量化等预处理。（3）根据主题一致性确定主题数量，构建 DTM 模型，获取“主题—词矩阵”和“文档—主题矩阵”文件。（4）对各个主题的主体强度进行计算，分析，将主题强度的演化趋势和关键词演化路径可视化，进行对比研究。

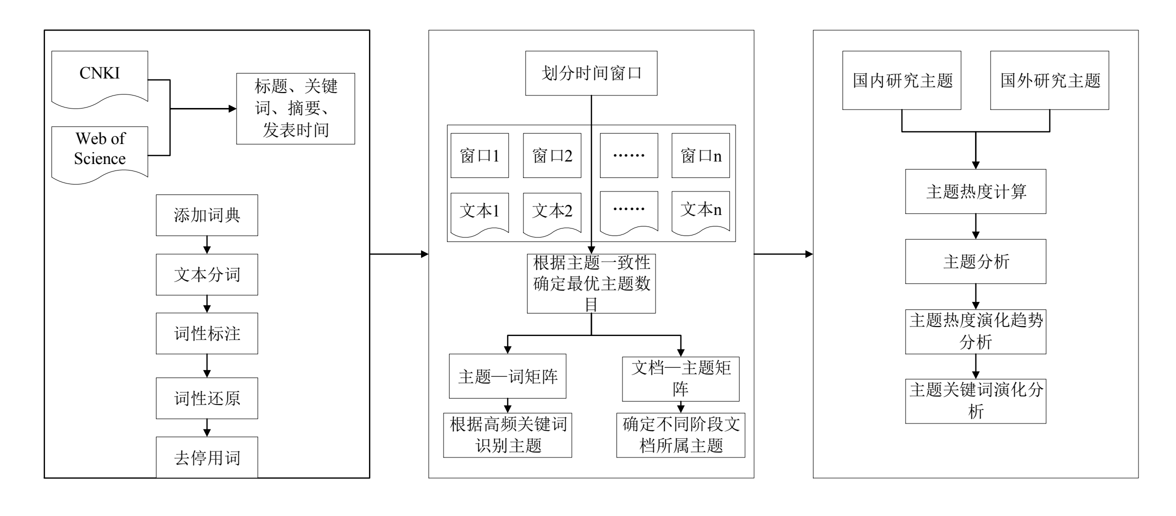


图 1 研究路线图

3.3 数据来源与处理

3.3.1 数据来源

本文分别以下载 2010 年至今 CNKI 和 Wos 政府数据相关文献，CNKI 的检索式为“主题=政府数据”，文献类型为 SCI、CSSCI 和北大核心，Wos 的检索式为“TS = government data”,文献类型为 Articles（检索时间为 2023 年 9 月 10 日）。对采集的文章进行人工筛选，去除会议记录，培训公告，序言和无摘要期刊等不相关内容后，共得到中文文献 646 篇，英文文献 888 篇。

3.3.2 划分时间切片

政府数据研究的热潮源起于 2009 年美国颁布的《开放政府指令》，根据发文数量、研究内容、研究方法等特征，将政府数据的研究历程划分为起步探索、快速发展和稳定推进三个阶段[18]。同时统计各年度发文量得图 3。



图 3 国内外各年度发文量

由图 3 可知 2010-2014 年，该阶段学者刚刚开始关注政府数据研究议题，年均发文量较低，所以将此阶段设定为起步探索阶段。2015-2018 年，随着政府数据的发展与建设，更多的学者开始投身于政府数据领域的研究中，本阶段政府数据相关的研究文献显著增加，所以将此阶段定义为快速发展阶段。2019 年至今，该阶段发文数量逐渐平稳，研究问题的专题深入已经进入相对成熟的阶段，将此阶段定义为稳定推进阶段。

3.3.3 语料库处理

将文献的摘要、标题、关键词和发表时间下载作为语料库。对于中文文献，调用 Python 中的中文分词工具 jieba 库得到每个文献的词集合，首先将文献中的关键词去重后添加到自定义用户词典中，使用哈工大中文停用词表，再根据模型效果对停用词表进行补充。对于英文单词，调用 NLTK 自然语言处理工具包，进行词性标注和词性还原后，将单词转换为小写，最后去除停用词。

3.3.4 主题个数确定

在进行 DTM 动态主题建模时，首先要通过计算主题一致性以获得最优主题个数，考虑到本研究语料库的数量，将主题个数设定为 2~20，主题一致性越大说明模型效果越好。由图 4 可知，根据计算结果，设置国内主题为 5，国外主题为 3 时，模型的收敛效果最好。

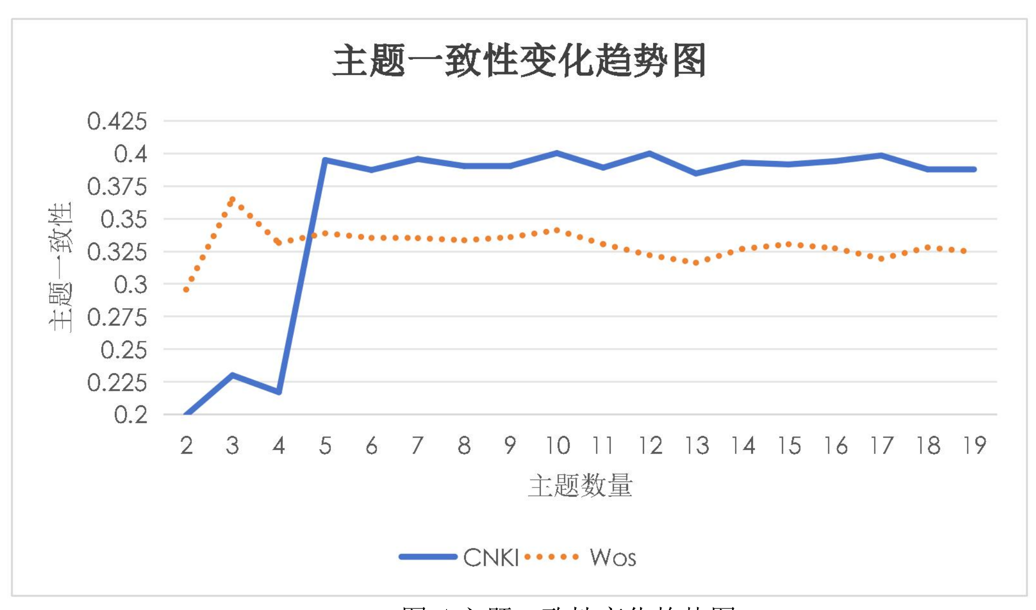


图 4 主题一致性变化趋势图

4 研究结果与分析

4.1 高频词分析

本研究对国内外语料库进行分词处理（剔除无意义词汇），然后进行词频统计选取高频词，利用 Python 中的 Wordcloud 工具包生成了高频词词云图，从整体上对国内外的研究方向进行分析，见图 5、图 6。





图 5　中文分词词云图　　　　图 6　英文分词词云图

通过对比国内外政府数据的词云图，可以看出，虽然国内外主要关注的重点都是政府数据政策和发展等相关的研究，但是通过价值、体系、发展等高频词，可以看出我国的政府数据相关研究主要角度是从政府自身出发看待政府数据开放或治理能产生的价值，而国外的高频词公众（Public）、使用（use）、服务（service）等从则表明从服务的角度对政府数据的开放、治理等政策进行研究，国内的政府数据研究更注重价值，而国外更注重服务。

4.2 主题研究

基于主题一致性的计算结果（见图 4），研究中将国内政府数据研究的主题分为 5 个，国外政府数据研究主题分为 3 个，见表 1，表 2。

表 1 国内政府数据研究主题

主题	关键词									
主题1：政府数据隐私	隐私	保护	个人隐私	风险	政策	美国	我国	数据开放	政府开放数据	政府数据开放共享
主题2：政府数据风险	风险	数据	政府	数据开放	价值	政策	研究	机制	影响	地方
主题3：政府数据政策	政策	政府数据开放平台	研究	政府数据治理	用户	价值	治理	开放	平台	数据
主题4：政府数据治理	治理	政府数据治理	评估	协同	政策	体系	OGD	数据治理	能力	隐私
主题5：政府数据评估	评估	政府数据治理	数据	政策	因素	影响	能力	治理	资源	中起

表 2 国外政府数据研究主题

主题	关键词									
主题1：政府数据质量	quality	public	research	portals	framework	service	local	transparency	model	citizens
主题2：政府数据信息	information	public	research	model	policy	quality	service	governments	state	local
主题3：政府数据评估	evaluation	public	information	countries	model	portals	quality	development	policy	increase

国内的研究主题类别高于国外，可以看出国内的学者对政府数据的研究具有较高的研究热情，并且研究的方向和范围相较于国外更为广泛。

4.2.1 国内研究主题分析

（1）政府数据隐私。政府数据隐私主要涉及的是政府数据开放过程中隐私

保护的问题，大部分学者将隐私保护和政策与法律关联在一起，黄如花等对我国[19]、美国[20]、英国[21]个人隐私保护的法律法规、政策和平台进行研究。刘凌等[22]从大数据的视角对国内外政府数据开放与个人隐私之间的关联进行研究，认为国内研究偏重于政策法规，理论研究较为稀缺，并提出了先导性模型以期打破当前困境。张晓娟等[23]以政府数据开放和个人隐私保护之间互相平衡为切入点，经过分析后指出了我国所面临的挑战并给出建议。

（2）政府数据风险。政府数据开放带来了经济价值和社会价值的同时，它所带来的风险也更加的复杂，政府数据的风险也是我国学者的研究重点之一。夏义堃[24]将政府数据开放风险分为：国家安全风险、政治风险、行政风险、隐私风险和经济风险，结合国内外的实践经历提出了导致政府数据开放风险的主要原因，认为应当建立相应的风险识别和评估机制，以提高风险应对能力。陈美等对荷兰[25]、日本[26]、西班牙[27]、巴西[28]的隐私风险控制进行研究分析各国对于风险防控的优势和劣势，进行归纳后为我国提供启示。赵需要针对政府数据开放中的国家秘密[29]、商业秘密[30]和个人隐私[31]泄露的风险进行调研，并提出了保护策略。

（3）政府数据政策。我国的政府数据政策的研究主要分为 3 个方面：数据保障政策、平台保障政策、隐私保护政策。黄如花、陈美、马海群等学者对此领域进行了深入的研究，目前我国政策主要存在体系不健全，监管不到位，各地政府建设差距过大等问题，还有许多改进的空间。

（4）政府数据治理。政府数据治理是当前国内外学术研究中比较热点的问题，同时也是推进数字化政府建设的首要前提。夏义堃等针对[32-34]政府数据治理的框架、内涵、生成背景进行了深入的研究与讨论，指出了我国政府数据治理面临的主要问题和困境，优化了治理的方式和路径。安小米等[35]分别对国内外政府数据治理现在进行分析，并总结出了我国政府数据治理的框架建设总体思路，从国家层面提出了新视角、新思路。

（5）政府数据评估。政府数据评估主题主要是对政府数据开放的效率和如何构建评估框架进行研究，张晓娟等[36]收集了 2012 年至 2017 年上海政府数据服务网全部的开放数据，采用更新积累度、更新及时度、更新增长度和更新转化度四个维度进行测算，创造了一系列完整的政府数据的评估体系。

4.2.2 国外研究主题分析

与国内研究相比，国外政府数据的研究主题相对较少，可以看出国外对于政府数据研究方向较为集中，并且国外政府数据的研究方向与国内存在差异，具体主题如下。

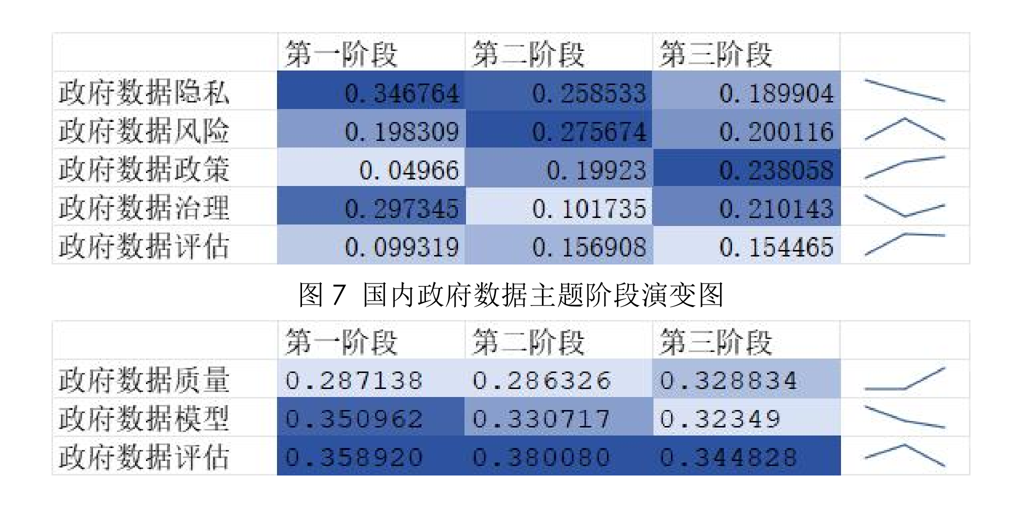
（1）政府数据质量。通过文献调研发现，在 2013 年之前国外主要从宏观层面对政府数据展开研究，包括但不限于政府数据开放能否支持民众参与决策、开放政府与政府数据之间的关系、政府数据重复利用的优势与建议等。在 2013 年后，国外的研究方向转向微观层面，对政府数据的质量、技术、应用等进行了深入研究。Oviedo 等[37]构建了政府数据开放门户中的数据质量模型，认为在政府数据开放的过程中，重复使用性和可分配性极大程度的影响了政府数据的质量，并且通过实验验证了其模型的准确性以及测量机制。Andrew Whitmore[38]指出政府数据的质量与用户利用数据的能力紧密相关，同时数据的质量也会影响数据提供者的备案质量，在研究中发现有关政府数据信息的主要问题表现在：信息准确性不足、信息缺乏和不完全。

（2）政府数据信息。通过关键词 information（信息）、policy（政策）、state（状态）等可以看出，国外对于政府数据的实施模型、政府数据政策和政府数据的相关状态较为关注，将此主题的主题标识拟定为政府数据信息。国外的研究中对概念模型的使用十分频繁，Hyoungjoo Park、Richard P. Smiraglia[39]利用概念参考模型（CIDOC CRM）对图书馆相关文件数据进行研究，以增强政府数据被发现的几率和重复利用的可能性。政策是一个研究领域的引导前提，Nugroho 等[40]认为政府开放数据的政策为政府数据的开放和使用提供了指导，但是一些国家因为对政府数据的政策不够完善，错失了很多从中获益的机会，于是对不同国家的数据政策进行了比较，并从中总结了经验和教训。

（3）政府数据评估。国外政府数据研究主要在质量、政策、基础设施等方面进行评估，以达到政府数据开发的效率最大化。Davide Ceolin 等[41]为了解决开放数据和未开放数据的差异和开放数据发展的趋势提出了两个评估开放政府的程序，以英国汉普郡警察局的数据为例，验证了这两种程序评估政府开放数据的真实性和可靠性。Charalampos Alexopoulos[42]等提出了一种评估开放政府数据系统的模型，根据模型中用户等级划分，从数据提供能力、下载能力、反馈能力等实现对政府数据系统进行改善和决策支撑。

4.3 主题强度演化分析

主题强度可以反映不同阶段的政府数据研究的战略重心，本文根据不同主题在不同阶段的主题强度绘制成热力图和折线图（见图 7、图 8），通过对主题强度和演变趋势折线图的变化进行分析，了解研究领域的状况，预测未来可能出现的热点主题。



	第一阶段	第二阶段	第三阶段	
政府数据隐私	0.346764	0.258533	0.189904	
政府数据风险	0.198309	0.275674	0.200116	
政府数据政策	0.04966	0.19923	0.238058	
政府数据治理	0.297345	0.101735	0.210143	
政府数据评估	0.099319	0.156908	0.154465	

图 7 国内政府数据主题阶段演变图

	第一阶段	第二阶段	第三阶段	
政府数据质量	0.287138	0.286326	0.328834	
政府数据模型	0.350962	0.330717	0.32349	
政府数据评估	0.358920	0.380080	0.344828	

图 8 国外政府数据主题阶段演变图

4.3.1 国内主题强度演化分析

由图 7 可知，国内研究中由两个主题呈上升趋势，分别为“政府数据政策”和“政府数据评估”；“政府数据隐私”呈持续下降趋势；“政府数据风险”和“政府数据治理”在研究过程中产生波动，但总体热度仍然较高。结合主题强度进行分析，可以看出在第一阶段中“政府数据隐私”和“政府数据治理”的主题强度是最高的，可以从中得出在政府数据研究的起步探索阶段我国比较注重个人隐私的保护和数据的治理，可能与中国以人为本的公共政策有关，在利用政府数据的同时我国会更加关注个人数据的安全。到了政府数据快速发展的第二阶段，“政府数据隐私”仍然是强度较高的主题，同时“政府数据风险”的相关研究得到了广大学者的关注，主题强度大幅提升，“政府数据治理”的关注度有所降低。到了稳定推进的第三阶段，可以从图中看出，强度最高的主题变为了“政府数据政策”，说明我国在 2019 年之后研究的重心主要放在了政策方面。同时从折线图可以看出，“政府数据隐私”方面的研究主题强度一直在持续下降，未来可能得到关注度仍然较小，“政府数据评估”和“政府数据政策”的热度在每个阶段均有提升，预测在未来关注度仍会持续上升。

4.3.2 国外主题强度演化分析

国外主题强度如图 8 所示，可以看出“政府数据评估”的热度在三个阶段均保持较高地位，“政府数据质量”相关研究在 2019 年后得到了大量关注，“政府数据模型”相关的研究的热度则一直持续降低。“政府数据质量”的相关研究在 2019 年之前发展十分稳定，但是在 2019 年后呈直线上升趋势，未来可能会持续上升发展为热点主题。与之相反的是“政府数据模型”相关的研究，一直处于持续下降的趋势，可能是因为在稳定推进的研究阶段，无论是相关的政策还是平台的研究，都已经较为成熟和完善，所以导致研究的热度降低。

4.4 主题关键词演化分析

在不同的时间阶段，同一个研究主题的重心也会存在差异。DTM 模型可以获取各个时间切片中的“主题-词矩阵”，通过分析主题在不同时间切片的研究重心的变化，以预测未来的研究内容，本文分别以国内持续上升的热点主题“政府数据政策”和国外主题强度最高的“政府数据评估”主题为例，输出主题关键词的演化路径，如图 9 和图 10 所示（图中展示前 10 个关键词）。

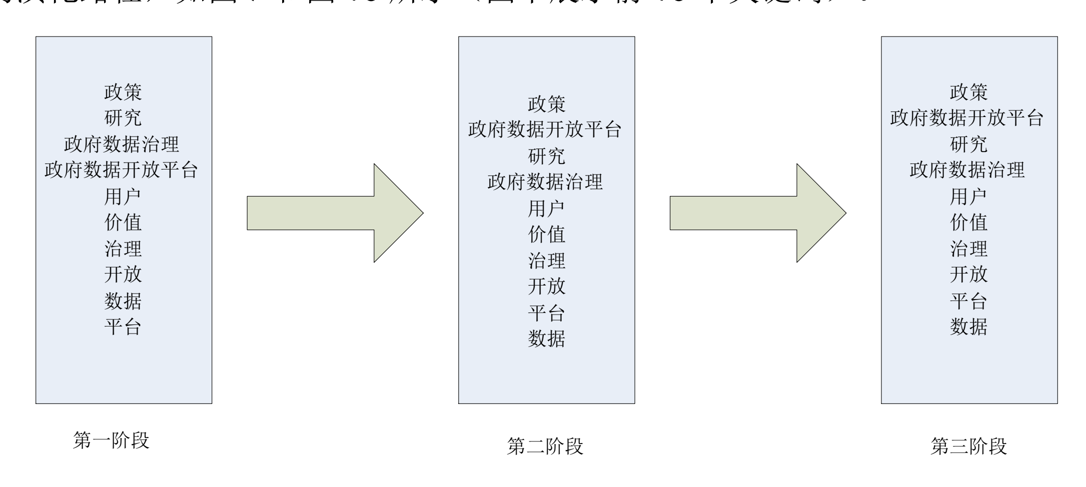


图 9 “政府数据政策”主题的关键词演化路径

4.4.1 国内主题关键词分析

通过图 9 可知在不同时间阶段下，“政府数据政策”主题下的核心关键词主要有“研究”、“政府数据治理”、“政府数据开放平台”、“用户”等。结合研究内容来看，我国的政策研究从数据的治理逐渐，像数据开放平台偏移，这也正对应着图 7 中，“政府数据治理”主题强度的下降。由此可见，政府数据开放平台的研究对于我国政府数据开放政策有着重要的研究意义，未来可以作为相关领域的研究重点。

4.4.2 国外主题关键词分析

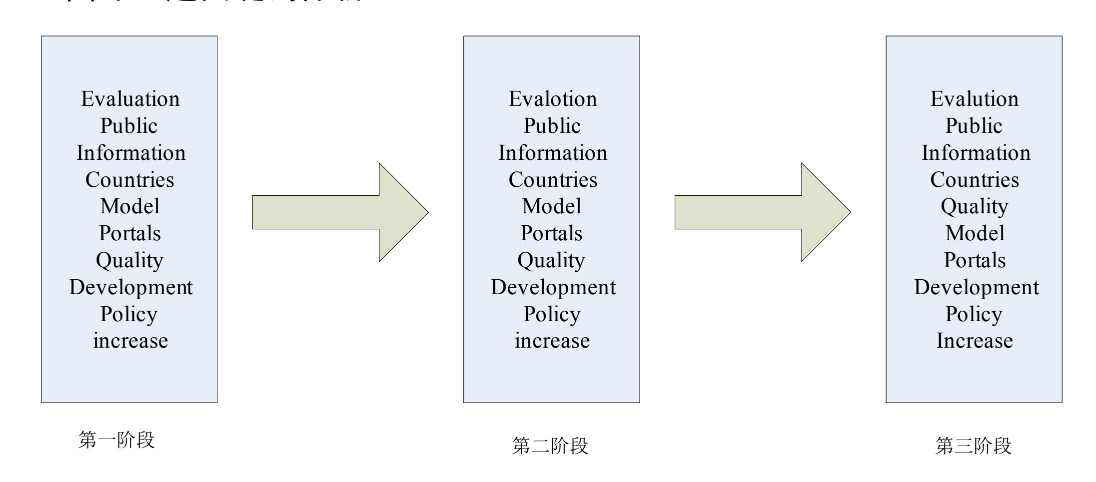


图 10 “政府数据评估”主题的关键词演化路径

由图 10 可知，“政府数据评估”主题处于不同发展阶段下相对稳定的核心关键词主要有“Public”、“Information”、“Countries”等。研究显示，在第一阶段和第二阶段中，“政府数据评估”主题的研究重心在于对公众、国家和政策的评价，到了第三阶段，数据的质量评估引起了学者的关注，主题强度得到了提升。结合具体关键词可知，国外的政府数据评估主要是为了评估政府、国家、门户网站发布的相关政策，作出改进，以提高数据质量，促进国家的发展。

5 结论和不足

本研究通过使用 DTM 动态主题模型对国内外政府数据相关论文进行了动态主题挖掘，并对结果进行了对比分析。总体来看，国内外政府数据研究在研究主题、主题强度及主题演化趋势等方面存在差异。主要表现在：①国内的研究主题数量较多，研究内容比较细化，对于个人隐私、数据治理、数据风险等问题较为关注，而国外的研究主题数量较少，主要对政府数据的质量、框架模型和评估方法进行研究。②国内的研究主题强度整体较低，起伏波动较大，国外的研究主题强度整体较高，主题演化趋势中较为稳定，波动较小，具有较高的传承性。③国外对于“政府数据评估”的相关研究较为重视，主题强度最高，国内关注度较低，虽然在最近有所提升，但与国外相比关注度仍然较低。通过对国内外政府数据的主题的对比分析，可以了解到不同时间切片国内外的研究重心和演变趋势，更好的找到我国的政府数据相关问题所在。同时本文存在一定不足：①文章的数据来源仅考虑了 CNKI 和 Wos 的研究论文，并未对国家政策文本进行分析，分析出

的结果可能存在一定误差。②文本语料中的用户词典采用文献关键词去重的方法构建，可能导致模型结果不够严谨和准确。

后续研究将会进一步扩充语料库的数量，在文献的基础上纳入政策文本，使生成的主题模型更加全面细化，对政府数据相关领域进行深入分析。

参考文献

[1]季统凯,刘甜甜,伍小强. 政府数据开放:概念辨析、价值与现状分析[J]. 北京工业大学学报,2017,43(03):327-334.

[2]张聪丛,郜颖颖,赵畅,杜洪涛. 开放政府数据共享与使用中的隐私保护问题研究——基于开放政府数据生命周期理论[J]. 电子政务,2018,(09):24-36.

[3]马海群,陶易. 基于 WSR 方法论的开放数据政策分析框架结构解析——以美国和加拿大为例[J]. 图书馆理论与实践,2018,(02):1-6.

[4]黄如花,温芳芳. 开放政府数据生命周期视角的我国政府数据资源管理政策文本内容分析——国家各部门的政策实践[J]. 图书馆,2018,(06):1-7+14.

[5]陈美. 开放政府数据隐私风险控制的先导性研究[J]. 情报理论与实践,2023,46(09):47-56.

[6]李晓彤,翟军,郑贵福. 我国地方政府开放数据的数据质量评价研究——以北京、广州和哈尔滨为例[J]. 情报杂志,2018,37(06):141-145.

[7]翟军,陶晨阳,李晓彤. 开放政府数据质量评估研究进展及启示[J]. 图书馆,2018,(12):74-79.

[8]孙小荃,翟军. 中美两国政府开放数据的元数据比较研究[J]. 图书馆杂志,2021,40(11):39-47.

[9]马海群,唐守利. 基于结构方程的政府开放数据网站服务质量评价研究[J]. 现代情报,2016,36(09):10-15+33.

[10]邹纯龙,马海群. 基于神经网络的政府开放数据网站评价研究——以美国 20 个政府开放数据网站为例[J]. 现代情报,2016,36(09):16-21.

[11]马海群,王今. 基于 DEA 的政府开放数据网站效率评价[J]. 数字图书馆论坛,2016,(06):2-7.

[12]张斌. 基于 LDA 主题模型的《国家情报法》法律本体分析[D].黑龙

江大学,2020.

[13]BLEI D M，Ng A Y，JOＲDAN M I． Latent dirichlet allocation ［J］． The journal of machine learning research，2003 (3) : 993 1022．

[14]BLEI D M，LAFFEＲTY J D． Dynamic topic models ［C］// Proceedings of the 23rd international conference on Machine learning．New York: ACM，2006: 113 － 120．

[15]邱均平,胡博,徐中阳,肖博轩. 基于DTM模型的国内外话语权研究主题挖掘及比较分析[J]. 情报理论与实践,2023,46(02):24-34.

[16]吴瑞鹏,李勇男,刘帅等.基于DTM的美国人工智能战略热点主题及演化分析[J/OL].情报杂志:1-10
[2023-10-10].http://hfffg5fce84748f1d4cc2hp5ofn6wvbwbc6wvu.fgfy.hlju.cwkeji.cn/kcms/detail/61.1167.g3.20230803.0749.016.html

[17]吴小兰,章成志. 基于DTM-LPA的突发事件话题演化方法研究——以H7N9微博为例[J]. 图书与情报,2015,(03):9-16.

[18]陈朝兵,张田. 2011-2020年国内外政府数据开放平台建设研究述评[J]. 图书情报知识,2022,39(01):130-140.

[19]黄如花,刘龙. 我国政府数据开放中的个人隐私保护问题与对策[J]. 图书馆,2017,(10):1-5.

[20]黄如花,李楠. 美国开放政府数据中的个人隐私保护研究[J]. 图书馆,2017,(06):19-24+76.

[21]黄如花,刘龙. 英国政府数据开放中的个人隐私保护研究[J]. 图书馆建设,2016,(12):47-52.

[22]刘凌,罗戎. 大数据视角下政府数据开放与个人隐私保护研究[J]. 情报科学,2017,35(02):112-118.

[23]张晓娟,王文强,唐长乐. 中美政府数据开放和个人隐私保护的政策法规研究[J]. 情报理论与实践,2016,39(01):38-43.

[24]夏义堃. 论政府数据开放风险与风险管理[J]. 情报学报,2017,36(01):18-27.

[25]陈美,梁乙凯. 荷兰开放政府数据的隐私风险控制研究[J]. 情报杂

志,2022,41(02):169-175+168.

[26]陈美,梁乙凯. 西班牙开放政府数据的隐私风险控制研究[J]. 情报杂志,2021,40(12):135-142.

[27]陈美,梁乙凯. 日本开放政府数据中的隐私风险控制研究[J]. 情报资料工作,2021,42(05):68-75.

[28]陈美,梁乙凯. 巴西开放政府数据的隐私风险控制研究[J]. 现代情报,2021,41(09):127-134.

[29]侯晓丽,彭靖,赵需要. 政府数据开放中国家秘密的泄露风险与保护策略[J]. 情报理论与实践,2018,41(07):53-59.

[30]赵需要,侯晓丽,彭靖. 政府数据开放中商业秘密的泄露风险与保护策略[J]. 情报理论与实践,2017,40(07):11-16.

[31]赵需要,彭靖. 政府数据开放中个人隐私的泄露风险与保护[J]. 信息安全研究,2016,2(09):792-801.

[32]夏义堃. 试论数据开放环境下的政府数据治理:概念框架与主要问题[J]. 图书情报知识,2018,(01):95-104.

[33]夏义堃. 试论政府数据治理的内涵、生成背景与主要问题[J]. 图书情报工作,2018,62(09):21-27.

[34]夏义堃. 政府数据治理的维度解析与路径优化[J]. 电子政务,2020,(07):43-54.

[35]安小米,王丽丽,许济沧,黄婕,白文琳. 我国政府数据治理与利用能力框架构建研究[J]. 图书情报知识,2021,38(05):34-47.

[36]唐长乐,张晓娟. 政府开放数据更新评估研究:类型、性能与测度[J]. 情报资料工作,2019,40(01):105-112.

[37]Oviedo E, Mazón J N, Zubcoff J J. Towards a data quality model for open data portals[C]//2013 XXXIX Latin American Computing Conference (CLEI). IEEE, 2013: 1-8.

[38]Whitmore A. Using open government data to predict war: A case study of data and systems challenges[J]. Government Information Quarterly, 2014, 31(4): 622-630.

[39]Park H, Smiraglia R P. Enhancing data curation of cultural heritage for

information sharing: A case study using open government data[C]//Metadata and Semantics Research: 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings 8. Springer International Publishing, 2014: 95-106.

[40]Nugroho R P, Zuiderwijk A, Janssen M, et al. A comparison of national open data policies: Lessons learned[J]. Transforming Government: People, Process and Policy, 2015, 9(3): 286-308.

[41]Ceolin D, Moreau L, O'Hara K, et al. Two procedures for analyzing the reliability of open government data[C]//Information Processing and Management of Uncertainty in Knowledge-Based Systems: 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part I 15. Springer International Publishing, 2014: 15-24.

[42]Alexopoulos C, Loukis E, Charalabidis Y. A methodology for determining the value generation mechanism and the improvement priorities of open government data systems[J]. Computer Science and Information Systems, 2016, 13(1): 237-258.

作者贡献说明：
马海群：负责论文的观点提炼，论文框架设计与修改。
孙正轩：负责相关资料收集，数据统计，论文撰写。

Thematic mining and comparative analysis of domestic and foreign government data research based on DTM model

Ma Haiqun[1] Sun Zhengxuan[2]

School of Information Management, Heilongjiang University, Harbin 150000

Abstract:[Purpose/Significance]To study the literature of domestic and foreign government data, analyze the differences in research themes and evolution trends at

home and abroad, which has important reference significance for understanding the current status of domestic and foreign government data research and promoting the innovation and development of domestic government data research. [Method/Process] The relevant literature of government data from 2010 to the present was selected, and the current situation and differences of domestic and foreign government data research were compared by constructing high-frequency word cloud maps, analyzing the evolution trend of theme intensity and the evolution path of theme keywords. [Results/Conclusions] There are great differences in government data-related research at home and abroad: (1) The number of domestic research topics is more large, and the research content is relatively detailed, while the number of foreign research topics is small, mainly on the quality, framework model and evaluation method of government data. (2) The overall intensity of domestic research themes is low, the fluctuations are large, and the overall intensity of foreign research themes is higher, and the theme evolution trend is relatively stable, the fluctuation is small, and it has high inheritance. (3) Foreign research focusing on "government data evaluation" has the highest thematic intensity and low domestic attention, although it has recently improved, but the attention is still low compared with foreign countries.
Keywords:government data; Open government data; DTM model; Theme mining; Comparative analysis

Classification number: G250